DORO: Distributional and Outlier Robust Optimization (Appendix)

Runtian Zhai [*1] Chen Dan [*1] J. Zico Kolter [1] Pradeep Ravikumar [1]

A. Proofs

A.1. Proof of Proposition 1

Since P_k is a mixture component of P with probability mass at least α, we can see that

$$\frac{dP_k}{dP} \leq \frac{1}{\alpha} \tag{19}$$

Notice that when $t \geq 1$,

$$f'_\beta(t) = \frac{1}{\beta - 1}(t^{\beta-1} - 1) > 0 \tag{20}$$

Hence, $f'_\beta(t)$ is an increasing function when $t > 1$, therefore,

$$\begin{aligned} D_\beta(P_k||P) &= \int f_\beta(\frac{dP_k}{dP})dP && (21)\\ &\leq \int f_\beta(\frac{1}{\alpha})dP && (22)\\ &= f_\beta(\frac{1}{\alpha}) && (23) \end{aligned}$$

Therefore, by the definition of β-DRO risk, we have completed the proof. □

A.2. Proof of Corollary 2

For any k, let $p_k = P(\mathcal{D}_k)$, then $P(z) = p_k P(z|\mathcal{D}_k) + (1 - p_k)P(z|\overline{\mathcal{D}_k})$ holds for all x. Let $Q = P_k$ and $Q'(z) = \frac{p_k - \alpha}{1-\alpha}P(z|\mathcal{D}_k) + \frac{1-p_k}{1-\alpha}P(z|\overline{\mathcal{D}_k})$. Then $P = \alpha Q + (1 - \alpha)Q'$, which implies that $\mathbb{E}_{P_k}[\ell(\theta; Z)] \leq \text{CVaR}_\alpha(\theta; P)$. Thus, $\mathcal{R}_{\max}(\theta; P) \leq \text{CVaR}_\alpha(\theta; P)$. On the other hand, for any Q such that there exists Q' satisfying $P = \alpha Q + (1 - \alpha)Q'$, we have $\frac{dQ}{dP}(z) \leq \frac{1}{\alpha}$ a.e., so that $D_{\chi^2}(Q \parallel P) \leq \frac{1}{2}(\frac{1}{\alpha} - 1)^2 = \rho$. Thus, $\text{CVaR}_\alpha(\theta; P) \leq \mathcal{R}_{D_{\chi^2},\rho}(\theta; P)$. □

A.3. Proposition 3

A.3.1. Proof of Proposition 3

By (6) and (11) we have

$$\begin{aligned} \mathcal{R}_{D_\beta,\rho,\epsilon}(\theta; P_{\text{train}}) &= \inf_{P'} \left\{ \mathcal{R}_{D_\beta,\rho}(\theta; P') : \exists \tilde{P}' \text{ s.t. } P_{\text{train}} = (1-\epsilon)P' + \epsilon\tilde{P}' \right\} \\ &= \inf_{P',\eta} \left\{ c_\beta(\rho)\mathbb{E}_{P'}[(\ell(\theta; Z) - \eta)_+^{\beta_*}]^{\frac{1}{\beta_*}} + \eta \right\} \\ &= \inf_{\eta} \left\{ c_\beta(\rho) \inf_{P'}\{[\int_{\mathbb{R}_+} P'((\ell(\theta; Z) - \eta)_+^{\beta_*} > u)du]^{\frac{1}{\beta_*}}\} + \eta \right\} \end{aligned} \tag{24}$$

By $P_{\text{train}} = (1 - \epsilon)P' + \epsilon\tilde{P}'$ we have for all ℓ_0,

$$P'(\ell(\theta; Z) \leq \ell_0) \leq \min \left\{ 1, \frac{1}{1-\epsilon} P_{\text{train}}(\ell(\theta; Z) \leq \ell_0) \right\} \tag{25}$$

and we can also show that there exists a $P^* = P'$ such that the equality is achieved in (25) for all ℓ_0: Since both ℓ and P_{train} are continuous, $P_{\text{train}}(\ell(\theta;z))$ is a continuous function of z for any fixed θ, so there exists an ℓ^* such that $P_{\text{train}}(\ell(\theta;Z) > \ell^*) = \epsilon$. Define

$$P^*(z) = \left\{ \begin{array}{cl} \frac{1}{1-\epsilon}P_{\text{train}}(z) & ,\ell(\theta;z) \le \ell^* \\ 0 & ,\ell(\theta;z) > \ell^* \end{array} \right. \tag{26}$$

For (26), we have $\int_{\mathcal{X}\times\mathcal{Y}} P^*(z)dz = \frac{1}{1-\epsilon}\int_{\ell(\theta;z)<\ell^*} P_{\text{train}}(z)dz = \frac{1}{1-\epsilon}P_{\text{train}}(\ell(\theta;Z) < \ell^*) = 1$ because $P_{\text{train}}(\ell(\theta;Z) = \ell^*) = 0$, so (26) is a distribution function.

Let $v = u^{\frac{1}{\beta_*}}$. Plugging $P^*(\ell(\theta;Z) \le \ell_0) = \min\left\{1, \frac{1}{1-\epsilon}P_{\text{train}}(\ell(\theta;Z) \le \ell_0)\right\}$ into (24) produces

$$\begin{aligned} \mathcal{R}_{D_\beta,\rho,\epsilon}(\theta;P_{\text{train}}) &= \inf_\eta \left\{ c_\beta(\rho)\left[\int_{\mathbb{R}_+}[1-P^*((\ell(\theta;Z)-\eta)_+^{\beta_*} \le v^{\beta_*})]dv^{\beta_*}\right]^{\frac{1}{\beta_*}} + \eta \right\} \\ &= \inf_\eta \left\{ c_\beta(\rho)\left[\int_{\mathbb{R}_+}[1-\frac{1}{1-\epsilon}P_{\text{train}}(\ell(\theta;Z) \le \eta + v)]_+ dv^{\beta_*}\right]^{\frac{1}{\beta_*}} + \eta \right\} \\ &= \inf_\eta \left\{ c_\beta(\rho)\left[\int_0^{(\ell^*-\eta)_+} \frac{1}{1-\epsilon}[(1-\epsilon) - P_{\text{train}}(\ell(\theta;Z) \le \eta + v)]_+ dv^{\beta_*}\right]^{\frac{1}{\beta_*}} + \eta \right\} \end{aligned} \tag{27}$$

On the other hand, we have

$$\begin{aligned} &\mathbb{E}_{P_{\text{train}}}[(\ell-\eta)_+^{\beta_*} \mid P_{Z'\sim P_{\text{train}}}(\ell(\theta;Z') > \ell(\theta;Z)) \ge \epsilon] \\ =&\frac{1}{1-\epsilon}\int_0^{\ell^*}(u-\eta)_+^{\beta_*} d(P_{\text{train}}(\ell \le u)) \\ =&\frac{1}{1-\epsilon}\left\{ \left[(u-\eta)_+^{\beta_*}P_{\text{train}}(\ell \le u)\right]_0^{\ell^*} - \int_0^{\ell^*} P_{\text{train}}(\ell \le u)d((u-\eta)_+^{\beta_*}) \right\} \\ =&\frac{1}{1-\epsilon}\left\{ (\ell^*-\eta)_+^{\beta_*}(1-\epsilon) - \int_0^{\ell^*} P_{\text{train}}(\ell \le u)d((u-\eta)_+^{\beta_*}) \right\} \\ =&\frac{1}{1-\epsilon}\left\{ \int_0^{(\ell^*-\eta)_+}(1-\epsilon)dv^{\beta_*} - \int_0^{(\ell^*-\eta)_+} P_{\text{train}}(\ell \le \eta + w)dw^{\beta_*} \right\} \end{aligned} \tag{28}$$

where $w = (u-\eta)_+$. Thus, (27) is equal to the right-hand side of (12). □

A.3.2. Extension to Arbitrary P_{train}

For any distribution P_{train}, we can obtain a similar but more complex formula (31). For any P_{train}, there exists an ℓ^* such that $P_{\text{train}}(\ell(\theta;Z) > \ell^*) \le \epsilon$ and $P_{\text{train}}(\ell(\theta;Z) < \ell^*) \le 1-\epsilon$. If $P_{\text{train}}(\ell(\theta;Z) = \ell^*) = 0$, then the proof above is still correct, so the formula is still (12).

Now assume that $P_{\text{train}}(\ell(\theta;Z) = \ell^*) > 0$. Similar to (26), define

$$P^*(z) = \left\{ \begin{array}{cl} \frac{1}{1-\epsilon}P_{\text{train}}(z) & ,\ell(\theta;z) < \ell^* \\ \left[1 - \frac{1}{1-\epsilon}P_{\text{train}}(\ell(\theta;Z) < \ell^*)\right]/P_{\text{train}}(\ell(\theta;Z) = \ell^*) & ,\ell(\theta;z) = \ell^* \\ 0 & ,\ell(\theta;z) > \ell^* \end{array} \right. \tag{29}$$

Then we still have $P^*(\ell(\theta;Z) \leq \ell_0) = \min\left\{1, \frac{1}{1-\epsilon} P_{\text{train}}(\ell(\theta;Z) \leq \ell_0)\right\}$, so (27) still holds. On the other hand, we have

$$\begin{aligned} &E_{P_{\text{train}}}[(\ell-\eta)_+^{\beta_*} \mid P_{Z'\sim P_{\text{train}}}(\ell(\theta;Z') > \ell(\theta;Z)) > \epsilon] \\ =&\frac{1}{P_{\text{train}}(\ell(\theta;Z) < \ell^*)} \left\{ \int_0^{(\ell^*-\eta)_+} (1-\epsilon) dv^{\beta_*} - \int_0^{(\ell^*-\eta)_+} P_{\text{train}}(\ell \leq \eta + w) dw^{\beta_*} \right\} \end{aligned} \tag{30}$$

Thus, the formula becomes

$$\begin{aligned} \mathcal{R}_{D_{\beta,\rho,\epsilon}}(\theta; P_{\text{train}}) = &\inf_\eta \{ c_\beta(\rho) (\frac{P_{\text{train}}(\ell < \ell^*)}{1-\epsilon} \mathbb{E}_Z[(\ell(\theta;Z) - \eta)_+^{\beta_*} \mid P_{Z'}(\ell(\theta;Z') > \ell(\theta;Z)) > \epsilon] \\ &+ \frac{1 - P_{\text{train}}(\ell < \ell^*)}{1-\epsilon} (\ell^* - \eta)_+^{\beta_*})^{\frac{1}{\beta_*}} + \eta \} \end{aligned} \tag{31}$$

A.4. Proofs of Results in Section 5

A.4.1. A Key Technical Lemma

The following lemma will be useful in the analysis of CVaR-DORO and χ^2-DORO: it controls the difference of dual objective in two distributions P, P' by their total variation distance, with the assumption that loss function l has bounded $2k$-th moment under P.

Lemma 8. *For any distributions P, P', non-negative loss function $l(\cdot, Z)$ and $1 \leq \beta_* < 2k$, such that $\mathbb{E}_P[l(\theta, Z)^{2k}] < \infty$, we have*

$$\mathbb{E}_P[(\ell-\eta)_+^{\beta_*}]^{\frac{1}{\beta_*}} \leq \mathbb{E}_{P'}[(\ell-\eta)_+^{\beta_*}]^{\frac{1}{\beta_*}} + \mathbb{E}_P[(l(\theta,Z)-\eta)_+^{2k}]^{\frac{1}{2k}} \text{TV}(P,P')^{(\frac{1}{\beta_*} - \frac{1}{2k})} \beta_*^{-\frac{1}{2k}} \cdot \left(\frac{2k}{2k-\beta_*} \right)^{\frac{1}{\beta_*}} \tag{32}$$

Proof. By the definition of total variation distance, we have

$$P(\ell(\theta;Z) > u) - P'(\ell(\theta;Z') > u) \leq \text{TV}(P,P') \tag{33}$$

holds for any $u \geq 0$.

By Markov's Inequality and the non-negativity of ℓ, we have for any $\eta \geq 0$,

$$P(\ell - \eta > u) \leq \frac{\mathbb{E}[(\ell-\eta)_+^{2k}]}{u^{2k}} := (\frac{s_{2k}}{u})^{2k} \tag{34}$$

where we introduced the shorthand $s_{2k} := \mathbb{E}[(\ell-\eta)_+^{2k}]^{\frac{1}{2k}}$ Using integration by parts, we can see that:

$$\mathbb{E}_P[(\ell-\eta)_+^{\beta_*}] = \int_\eta^\infty \beta_* (t-\eta)^{(\beta_*-1)} P(\ell \geq t) dt \tag{35}$$

$$= \int_0^\infty \beta_* u^{(\beta_*-1)} P(\ell - \eta \geq u) du \tag{36}$$

Thus,

$$\begin{aligned} \mathbb{E}_P[(\ell-\eta)_+^{\beta_*}] - \mathbb{E}_{P'}[(\ell-\eta)_+^{\beta_*}] &= \int_0^\infty \beta_* u^{(\beta_*-1)} \left(P(\ell-\eta \geq u) - P'(\ell-\eta \geq u) \right) du \\ &= \left(\int_0^M + \int_M^\infty \right) \left(\beta_* u^{(\beta_*-1)} \left(P(\ell-\eta \geq u) - P'(\ell-\eta \geq u) \right) du \right) \end{aligned} \tag{37}$$

Here, M is a positive parameter whose value will be determined later. Next, we will upper bound each of the two integrals separately. By equation 37,

$$\int_0^M \beta_* u^{(\beta_*-1)} \left(P(\ell-\eta \geq u) - P'(\ell-\eta \geq u) \right) du \leq \int_0^M \beta_* u^{(\beta_*-1)} \text{TV}(P,P') du \tag{38}$$

$$= M^{\beta_*} \text{TV}(P,P'), \tag{39}$$

which gives an upper bound for the first integral. For the second integral, notice that $P'(\ell-\eta \geq u)$ is non-negative and use equation 34, we have:

$$\begin{aligned}
\int_M^\infty \beta_* u^{(\beta_*-1)} \left(P(\ell-\eta\geq u) - P'(\ell-\eta\geq u)\right) du &\leq \int_M^\infty \beta_* u^{(\beta_*-1)} P(\ell-\eta\geq u) du && (40)\\
&\leq \int_M^\infty \beta_* u^{(\beta_*-1)} \left(\frac{s_{2k}}{u}\right)^{2k} && (41)\\
&= \frac{s_{2k}^{2k}}{2k-\beta_*}\cdot\frac{1}{M^{2k-\beta_*}} && (42)
\end{aligned}$$

Therefore, by setting $M = s_{2k}(\text{TV}(P,P')\beta_*)^{-1/2k}$ which minimizes the sum of two terms, we have

$$\mathbb{E}_P[(\ell-\eta)_+^{\beta_*}] - \mathbb{E}_{P'}[(\ell-\eta)_+^{\beta_*}] \leq \inf_{M>0}\left(M^{\beta_*}\text{TV}(P,P') + \frac{s_{2k}^{2k}}{2k-\beta_*}\cdot\frac{1}{M^{2k-\beta_*}}\right) = s_{2k}^{\beta_*}\text{TV}(P,P')^{1-\frac{\beta_*}{2k}}\beta_*^{-\frac{\beta_*}{2k}}\cdot\frac{2k}{2k-\beta_*} \tag{43}$$

Using the inequality $(A+B)^{\frac{1}{\beta_*}} \leq A^{\frac{1}{\beta_*}} + B^{\frac{1}{\beta_*}}$ when $\beta_* \geq 1$, we have:

$$\mathbb{E}_P[(\ell-\eta)_+^{\beta_*}]^{\frac{1}{\beta_*}} \leq \mathbb{E}_{P'}[(\ell-\eta)_+^{\beta_*}]^{\frac{1}{\beta_*}} + s_{2k}\text{TV}(P,P')^{\left(\frac{1}{\beta_*}-\frac{1}{2k}\right)}\beta_*^{-\frac{1}{2k}}\cdot\left(\frac{2k}{2k-\beta_*}\right)^{\frac{1}{\beta_*}} \tag{44}$$

□

A.4.2. Proof of Lemma 4

For any P' such that $P_{\text{train}} = (1-\epsilon)P' + \epsilon\tilde{P}'$ for some $\tilde{P}'$, let $U = P \wedge P'$, i.e. $U(z) = \min\{P(z), P'(z)\}$ for any $z \in \mathcal{X}\times\mathcal{Y}$. We have

$$(1-\epsilon)U(z) + \epsilon\tilde{P}(z) + \epsilon\tilde{P}'(z) \geq P_{\text{train}}(z) \quad \text{for any } z\in\mathcal{X}\times\mathcal{Y} \tag{45}$$

because both $\tilde{P}(z)$ and $\tilde{P}'(z)$ are non-negative. Integrating both sides produces

$$\int_{\mathcal{X}\times\mathcal{Y}} U(z)dz \geq \frac{1-2\epsilon}{1-\epsilon} \tag{46}$$

which implies $\text{TV}(P,P') \leq \frac{\epsilon}{1-\epsilon}$. Thus,

$$\mathcal{R}_{D,\rho}(\theta;P') \geq \inf_{P''}\{\mathcal{R}_{D,\rho}(\theta,P'') : \text{TV}(P,P'') \leq \frac{\epsilon}{1-\epsilon}\} \tag{47}$$

which together with (11) proves (13). □

A.4.3. Proof of Theorem 5, Analysis of CVaR-DORO

Proof of Theorem 5, CVaR-DORO. For any θ, by Lemma 4 we have

$$\text{CVaR}_{\alpha,\epsilon}(\theta;P_{\text{train}}) \geq \text{CVaR}_{\alpha,\epsilon}(\hat{\theta};P_{\text{train}}) \geq \inf_{P'}\{\text{CVaR}_\alpha(\hat{\theta};P') : \text{TV}(P,P') \leq \frac{\epsilon}{1-\epsilon}\} \tag{48}$$

By Lemma 8, we have for any $\eta \geq 0$ and $\text{TV}(P,P') \leq \frac{\epsilon}{1-\epsilon}$,

$$\begin{aligned}
\mathbb{E}_P[(\ell-\eta)_+] - \mathbb{E}_{P'}[(\ell-\eta)_+] &\leq \left(1+\frac{1}{2k-1}\right)\mathbb{E}_P[(\ell-\eta)_+^{2k}]^{\frac{1}{2k}}\text{TV}(P,P')^{1-\frac{1}{2k}}\\
&\leq \left(1+\frac{1}{2k-1}\right)\sigma_{2k}\text{TV}(P,P')^{1-\frac{1}{2k}}
\end{aligned} \tag{49}$$

Here, we used the fact that $0 \leq (\ell-\eta)_+^{2k} \leq \ell^{2k}$ Moreover, for any $\eta < 0$, $\mathbb{E}_P[(\ell-\eta)_+] - \mathbb{E}_{P'}[(\ell-\eta)_+] = \mathbb{E}_P[(\ell-0)_+] - \mathbb{E}_{P'}[(\ell-0)_+]$ because ℓ is non-negative. So (49) holds for all $\eta\in\mathbb{R}$. Thus, by (7) we have for any $\eta\in\mathbb{R}$,

$$\text{CVaR}_\alpha(\hat{\theta};P) \leq \alpha^{-1}\mathbb{E}_P[(\ell-\eta)_+] + \eta \leq \alpha^{-1}\left\{\mathbb{E}_{P'}[(\ell-\eta)_+] + \left(1+\frac{1}{2k-1}\right)\left(\frac{\epsilon}{1-\epsilon}\right)^{1-\frac{1}{2k}}\right\} + \eta \tag{50}$$

and taking the infimum over η, we have the following inequality holds for any θ:

$$\mathrm{CVaR}_\alpha(\hat{\theta}; P) \le \mathrm{CVaR}_\alpha(\hat{\theta}; P') + \left(1 + \frac{1}{2k-1}\right) \alpha^{-1} \sigma_{2k} \left(\frac{\epsilon}{1-\epsilon}\right)^{1-\frac{1}{2k}} \tag{51}$$

By (11) we have $\mathrm{CVaR}_{\alpha,\epsilon}(\theta; P_{\text{train}}) \le \mathrm{CVaR}_\alpha(\theta; P)$. Thus, by (48), taking the infimum over P' yields

$$\mathrm{CVaR}_\alpha(\hat{\theta}; P) \le \mathrm{CVaR}_{\alpha,\epsilon}(\theta; P_{\text{train}}) + \left(1 + \frac{1}{2k-1}\right) \alpha^{-1} \sigma_{2k} \left(\frac{\epsilon}{1-\epsilon}\right)^{1-\frac{1}{2k}} \tag{52}$$

$$\le \mathrm{CVaR}_\alpha(\theta; P) + \left(1 + \frac{1}{2k-1}\right) \alpha^{-1} \sigma_{2k} \left(\frac{\epsilon}{1-\epsilon}\right)^{1-\frac{1}{2k}} \tag{53}$$

Taking the infimum over θ completes the proof. □

A.4.4. Proof of Theorem 5, Analysis of χ^2-DORO

We begin with a structral lemma about the optimal dual variable η in the dual formulation equation 5. Recall that $\beta = \beta_* = 2$ for χ^2 divergence.

Lemma 9. *Let $\eta^*(P)$ be the minimizer of equation 5. We have the following characterization about $\eta^*(P)$:*

1. *When $\rho \le \frac{\mathrm{Var}_P[l(\theta,Z)]}{2\mathbb{E}[l(\theta,Z)]^2}$, we have $\eta^* \le 0$;*
   Furthermore, the DRO risk and optimal dual variable η^* can be formulated as:

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta; P) = \mathbb{E}_P[l(\theta, Z)] + \sqrt{2\rho \mathrm{Var}_P[l(\theta, Z)]} \tag{54}$$

$$\eta^* = \mathbb{E}_P[l(\theta, Z)] - \sqrt{\frac{\mathrm{Var}_P[l(\theta, Z)]}{2\rho}} \tag{55}$$

2. *When $\rho \ge \frac{\mathrm{Var}_P[l(\theta,Z)]}{2\mathbb{E}[l(\theta,Z)]^2}$, we have $\eta^* \ge 0$.*

Proof. (1) We will prove that for any $\rho > 0$,

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta; P) \le \mathbb{E}_P[l(\theta, Z)] + \sqrt{2\rho \mathrm{Var}_P[l(\theta, Z)]} \tag{56}$$

and the equality is achievable when $\rho \le \frac{\mathrm{Var}_P[l(\theta,Z)]}{2\mathbb{E}[l(\theta,Z)]^2}$.

By the definition of χ^2-DRO risk,

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta; P) = \sup_{Q: D_{\chi^2}(Q||P) \le \rho} E_Q[l(\theta, Z)] \tag{57}$$

Let $\mu := \mathbb{E}_P[l(\theta, Z)]$, notice that

$$\mathbb{E}_Q[l(\theta, Z)] = \mathbb{E}_P[l(\theta, Z)\frac{dQ}{dP}] \tag{58}$$

$$= \mathbb{E}_P[l(\theta, Z)] + \mathbb{E}_P[l(\theta, Z)\left(\frac{dQ}{dP} - 1\right)] \tag{59}$$

$$= \mu + \mathbb{E}_P[(l(\theta, Z) - \mu)\left(\frac{dQ}{dP} - 1\right)] \tag{60}$$

where in the last step we used the fact that $E_P \frac{dQ}{dP} = 1$.

By the definition of χ^2 divergence,

$$\mathbb{E}_P[\left(\frac{dQ}{dP} - 1\right)^2] = 2D_{\chi^2}(Q||P) \le 2\rho, \tag{61}$$

Therefore, by Cauchy-Schwarz inequality,

$$\mathbb{E}_P[(l(\theta, Z) - \mu)\left(\frac{dQ}{dP} - 1\right)] \leq (\mathbb{E}_P[(l(\theta, Z) - \mu)])^{1/2}\left(\mathbb{E}_P[\left(\frac{dQ}{dP} - 1\right)^2]\right)^{1/2} \tag{62}$$

$$\leq \sqrt{\mathrm{Var}_P[l(\theta, Z)] \cdot 2\rho} \tag{63}$$

Plug in this upper bound to equation 58 completes the proof of equation 56.

To see that the equality can be achieved when $\rho \leq \frac{\mathrm{Var}_P[l(\theta,Z)]}{\mathbb{E}[l(\theta,Z)]^2}$, we only need to verify that $\eta = \eta^*$ gives the same dual objective $\mathbb{E}_P[l(\theta, Z)] + \sqrt{2\rho\mathrm{Var}_P[l(\theta, Z)]}$. Since $\eta^* < 0$, we have

$$\mathbb{E}_P[(l(\theta, Z) - \eta^*)_+^2] \tag{64}$$

$$= \mathbb{E}_P[(l(\theta, Z) - \eta^*)^2] \tag{65}$$

$$= \mathbb{E}_P[(l(\theta, Z) - \mathbb{E}_P[l(\theta, Z)] + \sqrt{\frac{1}{2\rho}\mathrm{Var}_P[l(\theta, Z)]})^2] \tag{66}$$

$$= \mathbb{E}_P[(l(\theta, Z) - \mathbb{E}_P[l(\theta, Z)])^2] + 2\sqrt{\frac{1}{2\rho}\mathrm{Var}_P[l(\theta, Z)]}\mathbb{E}[(l(\theta, Z) - \mathbb{E}_P[l(\theta, Z)])] + \frac{1}{2\rho}\mathrm{Var}_P[l(\theta, Z)] \tag{67}$$

$$= \mathrm{Var}_P[l(\theta, Z)] + 0 + + \frac{1}{2\rho}\mathrm{Var}_P[l(\theta, Z)] = \frac{1 + 2\rho}{2\rho}\mathrm{Var}_P[l(\theta, Z)] \tag{68}$$

Therefore,

$$\sqrt{1 + 2\rho}\left(\mathbb{E}_P[(l(\theta, Z) - \eta^*)_+^2]\right)^{1/2} + \eta^* = \frac{1 + 2\rho}{\sqrt{2\rho}}\sqrt{\mathrm{Var}_P[l(\theta, Z)]} + \mathbb{E}_P[l(\theta, Z)] - \frac{1}{\sqrt{2\rho}}\sqrt{\mathrm{Var}_P[l(\theta, Z)]} \tag{69}$$

$$= \mathbb{E}_P[l(\theta, Z)] + \sqrt{2\rho\mathrm{Var}_P[l(\theta, Z)]} \tag{70}$$

and we have completed the proof.

(2) Let $g(\eta, P) = \sqrt{1 + 2\rho}\left(\mathbb{E}_P[(l(\theta, Z) - \eta)_+^2]\right)^{\frac{1}{2}} + \eta$ and recall that $\mathcal{R}_{D_{\chi^2},\rho}(\theta; P) = \inf_{\eta \in \mathbb{R}} g(\eta, P)$. To show that $\eta^* \leq 0$, we only need to prove that $g(\eta) \leq g(0)$ whenever $\eta < 0$, which is equivalent to:

$$\sqrt{1 + 2\rho}\left(\mathbb{E}_P[(l(\theta, Z) - \eta)^2]\right)^{\frac{1}{2}} \geq \sqrt{1 + 2\rho}\left(\mathbb{E}_P[l(\theta, Z)^2]\right)^{\frac{1}{2}} - \eta \tag{71}$$

Since both sides are non-negative, this inequality is equivalent to:

$$(1 + 2\rho)\mathbb{E}_P[(l(\theta, Z) - \eta)^2 \geq (1 + 2\rho)\mathbb{E}_P[l(\theta, Z)^2 + \eta^2 - 2\eta\sqrt{1 + 2\rho}\left(\mathbb{E}_P[l(\theta, Z)^2]\right)^{\frac{1}{2}} \tag{72}$$

After re-organizing terms, it remains to prove

$$2\rho\eta^2 - 2(1 + 2\rho)\eta E_P[l(\theta, Z)] + 2\eta\sqrt{1 + 2\rho}\left(\mathbb{E}_P[l(\theta, Z)^2]\right)^{\frac{1}{2}} \geq 0 \tag{73}$$

Since $\rho \geq \frac{\mathrm{Var}_P[l(\theta,Z)]}{2\mathbb{E}[l(\theta,Z)]^2}$, we have $(1 + 2\rho) \geq \frac{\mathbb{E}[l(\theta,Z)^2]}{\mathbb{E}[l(\theta,Z)]^2}$. Therefore,

$$LHS \geq 2\eta\sqrt{1 + 2\rho}\left(\mathbb{E}_P[l(\theta, Z)^2]\right)^{\frac{1}{2}} - 2(1 + 2\rho)\eta E_P[l(\theta, Z)] \tag{74}$$

$$= 2\eta\sqrt{1 + 2\rho}\left(\left(\mathbb{E}_P[l(\theta, Z)^2]\right)^{\frac{1}{2}} - \sqrt{1 + 2\rho}E_P[l(\theta, Z)]\right) \tag{75}$$

$$\geq 0 \tag{76}$$

where in the last step we used the assumption that $\eta \leq 0$. Therefore we have completed the proof.

□

Having prepared with Lemma 9, we are now ready to prove the χ^2-DORO part of Theorem 5.

Proof of Theorem 5, χ^2-DORO. We will first show that

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P) \le \mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P') + \sqrt{1+2\rho}(1+C_\rho)\sigma_{2k}\mathrm{TV}(P,P')^{\left(\frac{1}{2}-\frac{1}{2k}\right)}2^{-\frac{1}{2k}} \cdot \left(\frac{k}{k-1}\right)^{\frac{1}{2}} \tag{77}$$

This inequality will be proved by combining two different strategies: when $\eta^*(P')$ is relatively large, we will use an argument based on Lemma 8, similar to what we did in the analysis of CVaR-DORO. Otherwise, when $\eta^*(P')$ is small, we need a different proof which builds upon the structral result Lemma 9.

Define $C_\rho = \frac{\sqrt{1+2\rho}}{2\rho}$. Below we discuss two cases: $\eta^*(P') < -C_\rho\sigma_{2k}$ and $\eta^*(P') \ge -C_\rho\sigma_{2k}$.

Case 1: $\eta^*(P') < -C_\rho\sigma_{2k}$**.** When $\eta^*(P') < -C_\rho\sigma_{2k}$, by Lemma 9, we have

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P') = \mathbb{E}_{P'}[l(\hat{\theta},Z)] + \sqrt{2\rho\mathrm{Var}_{P'}[l(\hat{\theta},Z)]} \tag{78}$$

$$\eta^*(P') = \mathbb{E}_{P'}[l(\hat{\theta},Z)] - \sqrt{\frac{\mathrm{Var}_{P'}[l(\hat{\theta},Z)]}{2\rho}} < -C_\rho\sigma_{2k} \tag{79}$$

Therefore, we can lower bound $\sqrt{\mathrm{Var}_{P'}[l(\hat{\theta},Z)]}$ as:

$$\sqrt{\mathrm{Var}_{P'}[l(\hat{\theta},Z)]} \ge \sqrt{2\rho}\mathbb{E}_{P'}[l(\hat{\theta},Z)] + \sqrt{2\rho}C_\rho\sigma_{2k} \ge \sqrt{2\rho}C_\rho\sigma_{2k}, \tag{80}$$

and consequently, we have a lower bound for $\mathcal{R}_{D_{\chi^2},\rho}(\theta;P')$:

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P') = \mathbb{E}_{P'}[l(\hat{\theta},Z)] + \sqrt{2\rho\mathrm{Var}_{P'}[l(\hat{\theta},Z)]} \tag{81}$$

$$\ge \sqrt{2\rho\mathrm{Var}_{P'}[l(\hat{\theta},Z)]} \ge 2\rho C_\rho\sigma_{2k} = \sqrt{1+2\rho}\sigma_{2k} \tag{82}$$

On the other hand, by setting the dual variable $\eta = 0$, we have a simple upper bound for $\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P)$:

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P) \le \sqrt{1+2\rho}\mathbb{E}_P[l(\hat{\theta},Z)^2]^{1/2} \le \sqrt{1+2\rho}\sigma_{2k} \tag{83}$$

Combining equation 81 and equation 83, we conclude that $\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P') \ge \mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P)$ and the inequality is trivially true.

Case 2: $\eta^*(P') \ge -C_\rho\sigma_{2k}$ By Lemma 8, we have

$$\mathbb{E}_P[(\ell-\eta)_+^2]^{\frac{1}{2}} \le \mathbb{E}_{P'}[(\ell-\eta)_+^2]^{\frac{1}{2}} + \mathbb{E}_Z[(l(\theta,Z)-\eta)_+^{2k}]^{\frac{1}{2k}}\mathrm{TV}(P,P')^{\left(\frac{1}{2}-\frac{1}{2k}\right)}2^{-\frac{1}{2k}} \cdot \left(\frac{k}{k-1}\right)^{\frac{1}{2}} \tag{84}$$

holds for any $\eta \in \mathbb{R}$. Since $\eta^*(P') \ge -C_\rho\sigma_{2k}$, we can upper bound the $2k$-th moment $E_Z[(l(\theta,Z)-\eta^*(P'))_+^{2k}]^{\frac{1}{2k}}$ as:

$$E_Z[(l(\theta,Z)-\eta^*(P'))_+^{2k}]^{\frac{1}{2k}} \le E_Z[(l(\theta,Z)+C_\rho\sigma_{2k})_+^{2k}]^{\frac{1}{2k}} \tag{85}$$

$$\le E_Z[(l(\theta,Z)]^{\frac{1}{2k}} + C_\rho\sigma_{2k} = (1+C_\rho)\sigma_{2k} \tag{86}$$

Hence,

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P) \le \sqrt{1+2\rho}\mathbb{E}_P[(\ell-\eta^*(P'))_+^2]^{\frac{1}{2}} + \eta^*(P') \tag{87}$$

$$\le \sqrt{1+2\rho}\mathbb{E}_{P'}[(\ell-\eta^*(P'))_+^2]^{\frac{1}{2}} + \eta^*(P') + \sqrt{1+2\rho}(1+C_\rho)\sigma_{2k}\mathrm{TV}(P,P')^{\left(\frac{1}{2}-\frac{1}{2k}\right)}2^{-\frac{1}{2k}} \cdot \left(\frac{k}{k-1}\right)^{\frac{1}{2}} \tag{88}$$

$$= \mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta};P') + \sqrt{1+2\rho}(1+C_\rho)\sigma_{2k}\mathrm{TV}(P,P')^{\left(\frac{1}{2}-\frac{1}{2k}\right)}2^{-\frac{1}{2k}} \cdot \left(\frac{k}{k-1}\right)^{\frac{1}{2}} \tag{89}$$

Hence, we have proved the inequality equation 77. The rest of proof mimics CVaR-DORO. For any θ, by Lemma 4 we have

$$\mathcal{R}_{D_{\chi^2},\rho,\varepsilon}(\theta; P_{\text{train}}) \geq \mathcal{R}_{D_{\chi^2},\rho,\varepsilon} \geq \inf_{P'}\{\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta}; P') : \text{TV}(P, P') \leq \frac{\epsilon}{1-\epsilon}\} \tag{90}$$

By (11) we have $\mathcal{R}_{D_{\chi^2},\rho,\varepsilon}(\theta; P_{\text{train}}) \leq \mathcal{R}_{D_{\chi^2},\rho}(\theta; P)$. Thus, by (48), taking the infimum over P' yields

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta}; P) \leq \mathcal{R}_{D_{\chi^2},\rho,\epsilon}(\theta; P_{\text{train}}) + \sqrt{1+2\rho}(1+C_\rho)\sigma_{2k}\left(\frac{\epsilon}{1-\epsilon}\right)^{\left(\frac{1}{2}-\frac{1}{2k}\right)} 2^{-\frac{1}{2k}} \cdot \left(\frac{k}{k-1}\right)^{\frac{1}{2}} \tag{91}$$

$$\leq \mathcal{R}_{D_\beta,\rho}(\theta; P) + \sqrt{1+2\rho}(1+C_\rho)\sigma_{2k}\left(\frac{\epsilon}{1-\epsilon}\right)^{\left(\frac{1}{2}-\frac{1}{2k}\right)} 2^{-\frac{1}{2k}} \cdot \left(\frac{k}{k-1}\right)^{\frac{1}{2}} \tag{92}$$

Taking the infimum over θ completes the proof. □

A.4.5. PROOF OF THEOREM 6

We consider an optimization problem with the parameter space restricted to only two possible values $\Theta = \{\theta_0, \theta_1\}$. Our proof is constructive, which relies on the following distribution $P_{M,\Delta,\varepsilon}$:

$$l(\theta_0, Z) = 0, \quad l(\theta_1, Z) = \Delta \quad w.p. \quad (1-\varepsilon) \tag{93}$$

$$l(\theta_0, Z) = M, \quad l(\theta_1, Z) = \Delta \quad w.p. \quad \varepsilon \tag{94}$$

here M, Δ are some non-negative parameters whose value to be determined later and the probability is taken over the randomness of Z.

We have the following characterization of CVaR and χ^2-DRO risk:

Lemma 10 (DRO Risk of $P_{M,\Delta,\varepsilon}$)**.** *Assume that $\alpha \geq \varepsilon$ and $1 + 2\rho \leq \frac{1}{\varepsilon}$, we have the following closed-form expressions for CVaR and χ^2-DRO risk:*

$$\text{CVaR}_\alpha(\theta_0; P_{M,\Delta,\varepsilon}) = \frac{M\varepsilon}{\alpha} \tag{95}$$

$$\text{CVaR}_\alpha(\theta_1; P_{M,\Delta,\varepsilon}) = \Delta \tag{96}$$

and

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta_0; P_{M,\Delta,\varepsilon}) = M\varepsilon + M\sqrt{2\rho\varepsilon(1-\varepsilon)} \tag{97}$$

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta_1; P_{M,\Delta,\varepsilon}) = \Delta \tag{98}$$

Proof. Since $l(\theta_1, Z)$ is always a constant Δ, it's immediate to see $\text{CVaR}_\alpha(\theta_1; P_{M,\Delta,\varepsilon}) = \mathcal{R}_{D_{\chi^2},\rho}(\theta_1; P_{M,\Delta,\varepsilon}) = \Delta$. Hence we only need to focus on θ_0.

By the dual formulation of DRO risk, we have $\text{CVaR}_\alpha(\theta_0; P_{M,\Delta,\varepsilon}) = \inf_{\eta\in\mathbb{R}} h(\eta)$ and $\mathcal{R}_{D_{\chi^2},\rho}(\theta_0; P_{M,\Delta,\varepsilon}) = \inf_{\eta\in\mathbb{R}} g(\eta)$, where we use the shorthand $g(\eta)$ and $h(\eta)$ for

$$g(\eta) := \sqrt{1+2\rho}\left(\mathbb{E}_P[(l(\theta, Z) - \eta)_+^2]\right)^{\frac{1}{2}} + \eta \tag{99}$$

$$h(\eta) = \frac{1}{\alpha}\mathbb{E}_P[(l(\theta, Z) - \eta)_+] + \eta \tag{100}$$

Direct calculation gives:

$$g(\eta) = \begin{cases} \sqrt{1+2\rho}\sqrt{(\eta - \epsilon M)^2 + \varepsilon(1-\varepsilon)M^2} + \eta, & \text{for } \eta < 0 \\ \sqrt{\varepsilon(1+2\rho)}(M-\eta) + \eta, & \text{for } 0 \leq \eta \leq M \\ \eta, & \text{for } \eta > M \end{cases} \tag{101}$$

and

$$h(\eta) = \begin{cases} \frac{M\varepsilon - \eta}{\alpha} + \eta, & \text{for } \eta < 0 \\ \frac{\varepsilon(M-\eta)}{\alpha} + \eta, & \text{for } 0 \le \eta \le M \\ \eta, & \text{for } \eta > M \end{cases} \tag{102}$$

Therefore, when $\alpha \ge \varepsilon$ and $1 + 2\rho \le \frac{1}{\varepsilon}$, we have

$$\text{CVaR}_\alpha(\theta_0; P_{M,\Delta,\varepsilon}) = \inf_{\eta \in \mathbb{R}} h(\eta) \qquad = h(0) = \frac{M\varepsilon}{\alpha} \tag{103}$$

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta_0; P_{M,\Delta,\varepsilon}) = \inf_{\eta \in \mathbb{R}} g(\eta) \qquad = g(\varepsilon M - \frac{M\sqrt{\varepsilon(1-\varepsilon)}}{\sqrt{2\rho}}) = M\varepsilon + M\sqrt{2\rho\varepsilon(1-\varepsilon)} \tag{104}$$

and we have completed the proof. □

Equipped with Lemma 10, we are now ready to prove the main lower bound Theorem 6.

Proof of Theorem 6. Consider $P_{\text{train}} = P_{M,\Delta,\varepsilon}$. We have two different ways to decompose P_{train} into mixture of two distributions:

$$P_{\text{train}} = P_{M,\Delta,\varepsilon} = (1-\varepsilon)P_{M,\Delta,\varepsilon} + \varepsilon P_{M,\Delta,\varepsilon} = (1-\varepsilon)P_{0,\Delta,0} + \varepsilon P_{M,\Delta,0} \tag{105}$$

In other words, with only access to $P_{\text{train}} = P_{M,\Delta,\varepsilon}$, the learner cannot distinguish the following two possibilities:

- (a) The clean distribution is $P = P_{M,\Delta,\varepsilon}$, and the outlier distribution is $P' = P_{M,\Delta,\varepsilon}$.
- (b) The clean distribution is $Q = P_{0,\Delta,1}$, and the outlier distribution is $Q' = P_{M,\Delta,1}$.

Furthermore, as long as $M \le \sigma_{2k}\varepsilon^{-\frac{1}{2k}}$ and $\Delta \le \sigma_{2k}$, both P and Q satisfy the bounded $2k$-th moment condition $\mathbb{E}[l(\theta, Z)^{2k}] \le \sigma_{2k}^{2k}$. With our construction below, we can ensure that θ_1 is $\Theta(\Delta)$-suboptimal under P, while θ_0 is $\Theta(\Delta)$-suboptimal under Q. Therefore, in the worst case scenario, it's impossible for the learner to find a solution with $O(\Delta)$ sub-optimality gap under both P and Q.

CVaR lower bound Assume that $\alpha \ge \frac{1}{2}\varepsilon^{1-\frac{1}{2k}}$. Let $M = \sigma_{2k}\varepsilon^{-\frac{1}{2k}}, \Delta = \sigma_{2k}\frac{\varepsilon^{1-\frac{1}{2k}}}{2\alpha} \le \sigma_{2k}$. Recall that $P = P_{M,\Delta,\varepsilon}$, by Lemma 10, we have:

$$\text{CVaR}_\alpha(\theta_0; P) = \frac{M\varepsilon}{\alpha} = \frac{\sigma_{2k}}{\alpha}\varepsilon^{1-\frac{1}{2k}} = 2\Delta \tag{106}$$

$$\text{CVaR}_\alpha(\theta_1; P) = \Delta \tag{107}$$

Therefore,

$$\text{CVaR}_\alpha(\theta_0; P) - \inf_{\theta \in \Theta} \text{CVaR}_\alpha(\theta; P) = \Delta = \Omega(\frac{1}{\alpha}\sigma_{2k}\varepsilon^{1-\frac{1}{2k}}) \tag{108}$$

For $Q = P_{0,\Delta,1}$, both $l(\theta_0, Z)$ and $l(\theta_1, Z)$ are constants, and hence

$$\text{CVaR}_\alpha(\theta_0; Q) = 0 \tag{109}$$

$$\text{CVaR}_\alpha(\theta_1; Q) = \Delta \tag{110}$$

and

$$\text{CVaR}_\alpha(\theta_1; Q) - \inf_{\theta \in \Theta} \text{CVaR}_\alpha(\theta; Q) = \Delta = \Omega(\frac{1}{\alpha}\sigma_{2k}\varepsilon^{1-\frac{1}{2k}}) \tag{111}$$

Combining equation 108 and equation 111 completes the proof.

χ^2-DRO lower bound Assume that $\rho = O(\varepsilon^{\frac{1}{k}-1})$. Let $M = \sigma_{2k}\varepsilon^{-\frac{1}{2k}}$, $\Delta = \frac{M}{2}\left(\varepsilon + \sqrt{2\rho\varepsilon(1-\varepsilon)}\right) \le \sigma_{2k}$. Recall that $P = P_{M,\Delta,\varepsilon}$, by Lemma 10, we have:

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta_0; P) = M\varepsilon + M\sqrt{2\rho\varepsilon(1-\varepsilon)} = 2\Delta \tag{112}$$

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta_1; P) = \Delta \tag{113}$$

Therefore,

$$\mathcal{R}_{D_{\chi^2},\rho}(\theta_0; P) - \inf_{\theta\in\Theta} \mathcal{R}_{D_{\chi^2},\rho}(\theta; P) = \Delta = \Omega(\sigma_{2k}\sqrt{\rho}\varepsilon^{\frac{1}{2}-\frac{1}{2k}}) \tag{114}$$

For $Q = P_{0,\Delta,1}$, both $l(\theta_0, Z)$ and $l(\theta_1, Z)$ are constants, and hence

$$\text{CVaR}_\alpha(\theta_0; Q) = 0 \tag{115}$$

$$\text{CVaR}_\alpha(\theta_1; Q) = \Delta \tag{116}$$

and

$$\text{CVaR}_\alpha(\theta_0; Q) - \inf_{\theta\in\Theta} \text{CVaR}_\alpha(\theta; Q) = \Delta = \Omega(\sigma_{2k}\sqrt{\rho}\varepsilon^{\frac{1}{2}-\frac{1}{2k}}) \tag{117}$$

Combining equation 114 and equation 117 completes the proof.

□

A.4.6. Proof of Theorem 7

By Lemma 8, for any P' such that $\text{TV}(P, P') \le \frac{\epsilon}{1-\epsilon}$,

$$\text{CVaR}_\alpha(\theta; P) - \text{CVaR}_\alpha(\theta; P') \le 2\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}} \tag{118}$$

By Proposition 2, if $\mathcal{R}_{\max}(\theta; P) > 3\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}$, then $\text{CVaR}_\alpha(\theta; P) > 3\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}$, which implies that

$$\frac{\text{CVaR}_\alpha(\theta; P')}{\mathcal{R}_{\max}(\theta; P)} \ge \frac{\text{CVaR}_\alpha(\theta; P')}{\text{CVaR}_\alpha(\theta; P)} = 1 - \frac{\delta}{\text{CVaR}_\alpha(\theta; P)} \ge 1 - \frac{2\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}}{3\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}} = \frac{1}{3} \tag{119}$$

holds for any P' such that $\text{TV}(P, P') \le \frac{\epsilon}{1-\epsilon}$. By Lemma 4, taking the infimum over P' yields the first inequality of (18). Moreover, by Proposition 2, for any θ and P', $D_{\chi^2,\rho}(\theta; P') \ge \text{CVaR}_\alpha(\theta; P')$, which together with (119) yields the second inequality of (18). □

B. Experiment Details

B.1. Domain Definition

One important decision we need to make when we design a task with subpopulation shift is how to define the domains (subpopulations). We refer our readers to the Wilds paper (Koh et al., 2020), which discusses in detail the desiderata and considerations of domain definition, and defines 16 domains on the CivilComments-Wilds dataset which we use directly. The authors selected 8 features such as race, sex and religion, and crossed them with the two classes to define the 16 domains. Such a definition naturally covers class imbalance. There is no official domain definition on CelebA, so we define the domains on our own. Following their approach, on CelebA we also select 8 features and cross them with the two classes to compose the 16 domains. Our definition is inspired by (Sagawa et al., 2020), but we cover more types of subpopulation shift apart from demographic differences.

We select 8 features on CelebA: *Male*, *Female*, *Young*, *Old*, *Attractive*, *Not-attractive*, *Straight-hair* and *Wavy-hair*. We explain why we select these features as follows:

- The first four features cover sex and age, two protected features widely used in algorithmic fairness papers.

Table 1. Number of training instances in each domain of CelebA and CivilComments-Wilds.

CelebA	Blond	Others
Male	1387	66874
Female	22880	71629
Young	20230	106558
Old	4037	31945
Attractive	17008	66595
Not-attractive	7259	71908
Straight-hair	5178	28769
Wavy-hair	11342	40640
Total	162770	

CivilComments-Wilds	Toxic	Non-toxic
Male	4437	25373
Female	4962	31282
LGBTQ	2265	6155
Christian	2446	24292
Muslim	3125	10829
Other Religions	1003	5541
Black	3111	6785
White	4682	12016
Total	269038	

- We select the next two features in order to cover labeling biases, biases induced by the labelers into the dataset. Among the 40 features provided by CelebA, the *Attractive* feature is the most subjective one. Table 1 shows that among the people with blond hair, more than half are labeled *Attractive*; while among the other people, more than half are labeled *Not-attractive*. It might be that the labelers consider blond more attractive than other hair colors, or it might be that the labelers consider females more attractive than males, and it turns out that more females have blond hair than males in this dataset. Although the reason behind is unknown, we believe that these two features well represent the labeling biases in this dataset, and should be taken into consideration.

- We select the last two features in order to cover confounding variables, features the model uses to do classification that should have no correlation with the target by prior knowledge. Since the target is the hair color, a convolutional network trained on this dataset would focus on the hair of the person, so we conjecture that the output of the convolutional network is highly correlated with the hair style. In our experiments, we find that models trained with ERM misclassify about 20% of the test instances with blond straight hair, much more than the other three combinations.

Table 1 lists the number of training instances in each domain of each dataset. Each instance may belong to zero, one or more domains. In CivilComments-Wilds, the aggregated group size of the 16 groups is less than the total number 269,038, because most online comments do not contain sensitive words.

B.2. Model Selection

In Section 6 we assume access to a domain-aware validation set, which is not available in real domain-oblivious tasks. In this part we study several domain-oblivious model selection strategies, and discuss why model selection is hard.

We study the following model selection strategies:

- Max Average Accuracy: The model with the highest average accuracy in validation.

- Min CVaR: The model with the lowest CVaR risk ($\alpha = 0.2$) over the validation set.

- Min CVaR-DORO: The model with the lowest CVaR-DORO risk ($\alpha = 0.2, \epsilon = 0.005$) over the validation set.

Note that selecting the model that achieves the highest average accuracy over the worst α portion of the data is almost equivalent to the Max Average Accuracy strategy because the model with the highest average accuracy over the population also achieves the highest accuracy on the worst α portion (see e.g. (Hu et al., 2018), Theorem 1).

We conduct experiments on CelebA and report the results in Table 2. From the table we draw the following conclusions:

1. For every training algorithm, the oracle strategy achieves a much higher worst-case test accuracy than the other three strategies, and the gap between the oracle and the non-oracle strategies for DRO and DORO is larger than ERM. While it is expected that the oracle achieves a higher worst-case accuracy, the large gap indicates that there is still huge room for improvement.

Table 2. The average and worst-case test accuracies of the best models selected by different strategies. (%)

Training Algorithm	Model Selection	Average Accuracy	Worst-case Accuracy
ERM	Oracle	95.01 ± 0.38	53.94 ± 2.02
	Max Avg Acc	95.65 ± 0.05	45.83 ± 1.87
	Min CVaR	95.68 ± 0.04	44.83 ± 2.74
	Min CVaR-DORO	95.69 ± 0.04	44.50 ± 2.72
CVaR $(\alpha = 0.2)$	Oracle	95.52 ± 0.08	49.94 ± 3.36
	Max Avg Acc	95.74 ± 0.06	39.28 ± 3.58
	Min CVaR	95.79 ± 0.05	38.67 ± 2.06
	Min CVaR-DORO	95.81 ± 0.05	38.83 ± 2.05
CVaR-DORO $(\alpha = 0.2, \epsilon = 0.005)$	Oracle	92.91 ± 0.48	72.17 ± 3.14
	Max Avg Acc	95.60 ± 0.05	45.39 ± 3.22
	Min CVaR	95.58 ± 0.06	39.83 ± 2.37
	Min CVaR-DORO	95.56 ± 0.07	41.28 ± 3.26
χ^2-DRO $(\alpha = 0.2)$	Oracle	82.44 ± 1.22	63.36 ± 2.51
	Max Avg Acc	90.70 ± 0.26	20.67 ± 3.86
	Min CVaR	87.28 ± 2.05	21.44 ± 11.13
	Min CVaR-DORO	89.16 ± 1.41	25.50 ± 9.14
χ^2-DORO $(\alpha = 0.2, \epsilon = 0.005)$	Oracle	80.73 ± 1.41	65.36 ± 1.02
	Max Avg Acc	90.06 ± 0.57	22.06 ± 5.82
	Min CVaR	84.37 ± 4.08	29.83 ± 12.10
	Min CVaR-DORO	88.76 ± 0.81	23.61 ± 7.45

2. For χ^2-DRO/DORO, Min CVaR and Min CVaR-DORO work better than Max Average Accuracy. However, for the other three algorithms, Max Average Accuracy is better. This shows that model selection based on CVaR and selection based on CVaR-DORO are not good strategies.

3. With the three non-oracle strategies, ERM achieves the highest worst-case test accuracy. This does not mean that DRO and DORO are not as good as ERM, but suggests that we need other model selection strategies that work better with DRO and DORO.

The reason why Min CVaR is not a good strategy is that CVaR does not decrease monotonically with $\mathcal{R}_{\max}$. Corollary 2 only guarantees that CVaR is an upper bound of $\mathcal{R}_{\max}$, but the θ that achieves the minimum CVaR does not necessarily have the smallest $\mathcal{R}_{\max}$. For the same reason, Min CVaR-DORO is not a good strategy either.

Model selection under the domain oblivious setting is a very difficult task. In fact, Theorem 1 of (Hu et al., 2018) implies that no strategy can be provably better than Max Average Accuracy under the domain-oblivious setting, i.e. for any model selection strategy, there always exist $\mathcal{D}_1, \cdots, \mathcal{D}_K$ such that the model it selects is not better than the model selected by the Max Average Accuracy strategy. Thus, to design a provably model selection strategy, prior knowledge or reasonable assumptions on the domains are necessary.

B.3. Training Hyperparameters

On the COMPAS dataset, we use a two-layer feed-forward neural network activated by ReLU as the classification model. For optimization we use ASGD with learning rate 0.01. The batch size is 128. The hyperparameters we used in Table 2 were: $\alpha = 0.5$ for CVaR; $\alpha = 0.5, \epsilon = 0.2$ for CVaR-DORO; $\alpha = 0.5$ for χ^2-DRO; $\alpha = 0.5, \epsilon = 0.2$ for χ^2-DORO.

On the CelebA dataset, we use a standard ResNet18 as the classification model. For optimization we use momentum SGD with learning rate 0.001, momentum 0.9 and weight decay 0.001. The batch size is 400. The hyperparameters we used in Table 2 were: $\alpha = 0.1$ for CVaR; $\alpha = 0.2, \epsilon = 0.005$ for CVaR-DORO; $\alpha = 0.25$ for χ^2-DRO; $\alpha = 0.25, \epsilon = 0.01$ for χ^2-DORO.

On the CivilComments-Wilds dataset, we use a pretrained BERT-base-uncased model as the classification model. For optimization, we use AdamW with learning rate 0.00001 and weight decay 0.01. The batch size is 128. The hyperparameters we used in Table 2 were: $\alpha = 0.1$ for CVaR; $\alpha = 0.1$, $\epsilon = 0.01$ for CVaR-DORO; $\alpha = 0.2$ for χ^2-DRO; $\alpha = 0.2$, $\epsilon = 0.01$ for χ^2-DORO.

References

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.